Assessing statistical results in *MOR* articles:

An essay on verifiability and ways to enhance it

MING LI
University of Liverpool Management School
Chatham Street
Liverpool, UK
L69 7ZH
Ming.Li2@liverpool.ac.uk

BARTON M. SHARP
Department of Management
Northern Illinois University
Barsema Hall
DeKalb, IL 60115
bsharp1@niu.edu

DONALD D. BERGH
Daniels College of Business
The University of Denver
2101 S. University Blvd.
Denver, CO 80208
dbergh@du.edu

Assessing statistical results in *MOR* articles:

An essay on verifiability and ways to enhance it

Introduction

Readers of empirical research trust reported results and conclusions to be completely honest and accurate and place faith in the peer review process to have caught instances in which they were not. However, as human beings, researchers make mistakes. For example, a study reports that 18 percent of statistical results in a sample of 281 studies are incorrectly reported (Bakker & Wicherts, 2011). Similarly, surveys of psychologists and management researchers find that they, or colleagues, have engaged in questionable academic practices with respect to reporting empirical findings (Bedeian, Taylor, & Miller, 2010; John, Loewenstein, & Prelec, 2012). These observations have led to questions about validity of scientific claims. To address this criticism among several others, the editors of *Management Organization Review* (*MOR*) have taken the lead and specified future policies, and some of the policies seek to enhance data transparency and reporting practice (Lewin et al., forthcoming). However, the editorial stops short of identifying whether previous reporting practices in this journal might not sufficiently safeguard problems to creep in and authors can either blunder or even mislead their way to publication.

We propose that it is timely to examine the previous reporting practices and verify the accuracy of reported empirical findings in management and organization research. By definition, verification occurs when claim is confirmed or substantiated from its own evidence (www.dictionary.com). If 'the truth is under attack'(Levine, 2012), verification builds the first line of defence of credibility of our research. In this essay, we report attempts

to verify statistical results of a random sample of empirical articles appearing in *MOR*. Our purpose is to document how complete results and data are reported as well as develop an initial estimate of the accuracy of reported findings. We selected *MOR* because its leadership position in publishing standards and ethics are likely to shape the practices adopted by other journals as well. By assessing *MOR*, we seek to help editors gain insights of verifiability of their own published articles. More broadly, we seek to provoke debate about current reporting practices, guide the development of future reporting practices, and provide recommendations to enhance and protect the empirical foundation of management and organization research.

Method

Verification tests[1]

Verifying a study's empirical findings implies accessing and testing the original data (Bergh, Sharp, & Li, in press, 2017) However, several verification methods exist that do not require original data and can use descriptive and test statistics instead. After an extensive survey of literatures in management, psychology, economics, and sociology, Bergh and colleagues identified three such methods that can be used by independent parties to reproduce findings and verify their levels of accuracy and validity (Bergh et al., in press, 2017).

Test One examines the congruence of reported statistical results based on null hypothesis significance testing (NHST) with their statistical parameters. This test reproduces *p*-values based on reported test statistics and degrees of freedom (*df*), and compares them

[1] For more details of verification tests, please refer to Bergh, Sharp & Li (in press).

with reported *p*-values. The test requires disclosure of coefficients, *p*-values, standard errors (SEs) or parameter statistics (*t, f, z*), and the degrees of freedom (Bakker & Wicherts, 2011).

Test Two examines whether reported findings can be reproduced from disclosed data. Using the study's reported descriptive statistics, including variables' means, standard deviations (SDs), number of cases (Ns) and their correlations, these data are arranged into a matrix and inputted into a statistical software package instead of the original raw data. (Zientek & Thompson, 2009). To conduct this test, a correlation matrix is required, and it must include means, SDs, Ns and correlation coefficients of all variables included in tested models of a focal study.

Test Three uses a simulation-based procedure as if published research were repeated for numerous times (e.g., 1000 times) and each repetition drawn a new random observations from the same underlying population as the original research. The test estimates how many coefficients may be over- or under-stated relative to an expected effective size (see Goldfarb & King, 2016). The test results allow researchers to characterize the stability or generalizability of published findings. Similar to Test One, this test also requires disclosure of coefficients, *p*-values, standard errors (SEs) or parameter statistics (*t, f, z*), and the degrees of freedom.

Sample and data

We first selected a sample of *MOR* articles that reported Ordinary Least Squares (OLS) regression analysis, as this technique is widely used throughout management and organization research and can be considered within the context of all three verification tests.[2] To identify a random sample of articles, we identified the first article in each year's *MOR*

[2] Several analytical techniques can be tested using matrices of descriptive data including structural equations modeling, analysis of variance, discriminant analysis and factor analysis.

issues from year 2005 to 2015 that applied OLS regressions. This process led to a sample of ten articles[3]. Next, we examined each article to determine whether reporting practices permitted re-testing. We found three articles published in 2006, 2008 and 2010 which did not report enough details about their data to permit application of the tests. We then identified replacement articles by moving on to the second published article which contained OLS regressions appearing in each of these three years. Overall 13 articles are included in our sample and 10 permit application of at least one of the three tests.

The reliability of the data entry was tested through comparisons across study authors. One author conducted Test One using Excel software and Test Two by SPSS Matrix Data syntax[4]. Another author conducted Test Two by Stata corr2data command and Test Three using the Stata code provided by Goldfarb & King (2016) as an online supplement to their article. No discrepancies were identified across the two authors where the same analyses were conducted by both.

Results

Of the 13 articles in the sample, 6 could be verified using Test One and Test Three. The other 7 did not report sufficient data to permit re-resting, including standard errors (SEs) or *t* values. Test Two could be applied to all models in 3 out of the 13 articles (23%); and to some models in 5 articles (38.5%). Several models could not be retested due to their use of interaction terms that were not reported in correlation tables. Test Two could not be applied to 5 articles (38.5%), 4 articles due to missing correlation table, missing means, SDs, dummy variables in correlation table; and one article due to correlation matrix is not positive semi-definite.

[3] We screened through all published articles in 2012 and did not find any research articles apply OLS or regression method.
[4] Coding syntax and commands for all three tests are available upon request.

Findings from Test One

The 13 articles collectively report 475 coefficients that can be reproduced by Test One. 44 of the 475 (9%) reproduced *p*-values are different than those reported, indicating that 91% could be verified. Among the 44 *p*-values that are not congruent, 10 could be due to rounding error, 7 reproduced *p*-values are actually more significant than reported (5 of them are for intercepts). The remaining 27 reproduced *p*-values are less significant than reported *p*-values, and among them 13 are related to hypotheses testing. 12 of these 13 reproduced *p*-values are only at higher band of significance level from the reported *p*-values; hence do not influence the conclusions drawn from them. Only one (8%) of the 13 coefficients has error that could influence hypothesis conclusion.

Findings from Test Two

Test Two reproduced 243 coefficients. 39 (16%) of the reproduced *p*-values are different from reported *p*-values, 7 are intercepts coefficients and 5 reproduced *p*-values are about the same as the reported value so the difference could be due to rounding error. Eight are only difference in statistically accepted significant levels and do not influence conclusions; 20 have a reproduced significant *p*- value and non-significant reported *p*-value; 9 have a reproduced non-significant *p*-value and reported significant *p*- value. The collective verifiability rate of *p*- values is approximately 85 percent. 22 (9%) reproduced coefficients have a sign opposite of reported sign; hence the verifiability rate of coefficients signs is 91%.

Overall, 51 reproduced coefficients pertained to hypotheses. We find that 12 (23.5%) of the reproduced *p*-values are different than those reported and 6 have different signs. Among them, two *p*-values are different only in significant level and do not influence hypothesis conclusion; five have reproduced non-significant *p*-values and reported significant *p*-values hence their supported hypotheses are rejected in reproduction and five have

reproduced significant *p*-values and reported non-significant *p*-values hence their unsupported hypotheses are actually supported in reproduction. Only 2 of the 11 coefficients have a reproduced sign opposite to reported sign. Overall, about 20% of reproduced *p*-values that are pertained to hypotheses testing could revise the conclusions in three articles in our sample.

Findings from Test Three

Figure 1 presents the output of Test Three - the Goldfarb and King (2016) simulation procedure applied to 475 coefficients reported in the sample papers. The plus (+) and minus (-) signs on the graph represent the upper and lower 95% confidence intervals on the number of coefficients which would be expected to have a given *t*-value (represented by the categories on the horizontal axis) should the reported regressions be re-run 1000 times, with each re-run being conducted on a new random draw from the same underlying populations as those used in the original research. The gray bars represent the number of coefficients that were reported to have each *t*-value in the original research.

To interpret these results we specifically look for instances where particular *t*-values were either over- or under-represented in the original research compared to what we would expect to see over multiple iterations of the sampling procedure. We are especially sensitive to over-reporting of coefficients with t-values above 1.96 since that is the cut-off for the traditional $p<0.05$ level of significance. Over-reporting of significant *t*-values compared to what we would expect if the research were repeated could indicate an error of omission or commission on the part of the original authors. This might take the form of "cherry-picking" particular variables, regression models, or even entire papers where authors only submit results that indicate significance and leave insignificant findings in their desk drawer, or it could be the result of authors simply "shading" their results to the significant side.

In the current *MOR* sample we see a slight indication of such a problem. In the articles coded for our review, there were 29 coefficients reported to have t=2, just on the significant side of the *t*=1.96 cut-off. Based on the Goldfarb and King simulation procedure we would have expected to see no more than 16 coefficients with that particular *t*-value. This finding suggests that authors may have either altered results or selectively reported only those regressions which gave them the statistical significance that is so important for publication. Noticeably that calls into question the extent to which those findings would be robust in new samples from the same populations.

Discussion

Concerns about the state of scientific research are widespread, and editors are now adopting new reporting standards to protect the integrity of the research published in their journals. This essay examines the reporting practices and verifiability of empirical findings reported in *MOR* articles. Although our study draws from a small sample of 13 empirical articles, our findings reveal some reporting practices that permit verification while others may impede it. Statistical results that could be reproduced were typically verified at a level around 90 percent based on Test One and Test Two, though findings from our simulation based Test Three indicates that more coefficients were reported as statistically significant than should have been. Taken together, these results suggest that while the overall verifiability rate is relatively high, errors in findings appear to exist within *MOR*.

Reporting and Disclosure Practices

Any attempt to verify reported empirical findings is contingent upon researchers' disclosure of basic statistics. Three articles (23%) report all required statistics so could apply all three verification tests. Despite that we could verify, fully or partially, the findings of 10 out of the 13 articles (77%), many articles did not report statistics sufficiently to permit

verification. Test One and Test Three could be applied to slightly less than half of the articles, but could not be applied to slightly more than half of the articles due to missing standard errors (SEs) or *t* values. Test Two could be applied to 8 of the 13 articles, but could not be fully applied to more than two third of articles due to missing correlation tables, missing means, SDs, sample size values (Ns) in correlation tables, and missing descriptive statistics in correlation table for dummy control variables and interaction terms when authors tested moderation hypotheses.

If authors had reported required basic statistics, then all verification tests could have been applied. Our recommendations support those of Bergh and colleagues (in press; 2016) whom have called upon reviewers and editors to expand reporting of basic statistical figures. Firstly, it is essential for authors to report correlation tables if they conduct non-experimental quantitative research. The correlation table is the first point of reference for readers to gain understanding of a study's primary data (Bedeian, 2014), and verification Test Two is not possible without it. The 2010 *Publication Manual of the American Psychological Association* (APA) requires the disclosure of variable means, sample sizes, and variance—covariance (or correlation) matrix or matrices for multivariable analytic systems such as multivariate analyses of variance, regression analyses, structural equation modeling analyses, and hierarchical linear modeling (American Psychological Association, 2010). It is now time for management journals to strictly implement these requirements.

Secondly, we add our support to recent calls for authors to disclose all study variables (including control variables, product terms, dummy variables) and their respective means SDs, number of cases, estimated reliability scores if applicable, correlation coefficients and their significance levels (Bedeian, 2014; Bergh et al., in press, 2017). Reporting these statistics in full is not only important for readers to gain understanding of a study's primary data (Bedeian, 2014), but also enable them to conduct verification tests to examine the

validity of a focal study and then the replication studies that follow (Bergh et al., in press, 2017). Further, such figures are critical for effective meta-analyses and replications.

Verifiability

The overall verifiability rate of reported statistics published in *MOR* is 91% based on Test One, which is slightly better than 89% in medical journals (García-Berthou & Alcaraz, 2004), 86% in psychiatry journals (Berle & Starcevic, 2007) and 82% in psychology journals (Bakker & Wicherts, 2011). Also some errors may be due to reporting two decimal places rather than 3 decimal places, therefore the overall verifiability rate could be higher than 91%. The overall verifiability rate based on Test Two is about 85% for *p*-values and 91% for coefficients signs. Results of two tests indicate the overall the reporting verifiability in *MOR* journals is comparable with if not slightly better than other journals. However there is room to improve for the most desired 100% accuracy rate. The accuracy of reported findings is not only an indicator of a focal study's credibility, but also important for knowledge accumulation. For example, when *p*-values are used in Meta-analysis, the errors of reporting *p*-values could bias meta-analysis results considerably.

We are specifically interested in verifiability of coefficients for hypotheses testing due to the role of these coefficients in knowledge accumulation. Based on Test One, 13% of 100 coefficients for hypotheses testing have a higher reproduced p-values, potentially could influence conclusions drawn from them. A closer examination of these 13 coefficients found 12 of them are only at higher band of significance level from the reported *p*-values, hence do not influence the conclusions drawn from them. Only one of the 13 coefficients has error and could influence hypothesis conclusion, leaving 1 out of 13 articles (8%) may have a hypothesis not be supported. This, in comparison with that around 15% of the articles in psychology journals contained at least one statistical conclusion that proved, upon recalculation to be incorrect (Bakker & Wicherts, 2011), is better and encouraging.

The overall verifiability rate of Test Two is quite comparable Test One. However a closer examination of 51 reproduced coefficients that are for hypotheses testing reveals a lower verifiability rate of about 80% which is not as high as Test One. The 20% unverifiable *p*-values of hypotheses testing coefficients could due to reasons ranging from error or typographical mistake in the published tables of descriptive statistics or correlations or in the published regression results, to authors chose to falsify results by reporting a coefficient sign or *p*-value different than that which resulted from their regressions, to the regressions were run on a dataset that differed in some way from that described in the tables of means, standard deviations, and correlations, such as when an author might run regressions on a cherry-picked subsample of the original data in order to snoop for significant findings (Bergh et al., in press, 2017). No matter what reasons they are, the conclusions supported or rejected based on the reported *p*-values in three articles (23%) in the sample could be questioned. Further studies that build on these studies could also be questioned consequently. Our accumulative knowledge requires higher verifiability of reported findings.

Test Three similarly suggests a potential issue with the way data and results are being presented. The test results indicate possibilities that authors either report coefficients to be more significant than they really are, or are selective about which models to present. For science to be valid and meaningful we must be willing to fairly report all results, rather than only those which demonstrate statistical significance or support a hypothesis. Selective reporting gives a faulty impression of a generalizable phenomenon when the results were in fact an artefact of only samples and very specific models that were chosen.

Where do we go from here?

We recognize that the tests we adopt also do not apply to all statistical analyses. Yet, since OLS regression is the most trained rudimentary statistical analysis, there is no reason to

expect articles applying other analytical approaches would be more verifiable. Verification tests offer a mechanism for assessing the verifiability of a focal study. We call for verification tests as part of the review process, and for editors to verify reported findings when a manuscript reaches the conditional acceptance stage as a minimum effort (e.g., Bergh et al., in press; 2016).

It is noteworthy that *MOR* published articles had higher verifiability rates in comparison with journals in other disciplines. Still there is room to improve both in the practice and accuracy of reporting. It appears the practice of only reporting regression coefficients and *p*-values in asterisks format is still quite prevailing and missing report of SEs or *t*-values make verification of *p*-values impossible if there are missing means, SDs, Ns or variables in correlation matrix as well. Our reflection prompts us to ask two fundamental questions.

1. What reporting requirements should be followed?

"Uniform reporting standards make it easier to generalize across fields, to more fully understand the implications of individual studies, and to allow techniques of meta-analysis to proceed more efficiently" (American Psychological Association, 2010, p21). We advocate for following reporting requirements:

(1) Authors of nonexperimental studies that apply multivariate methods to include correlation matrices in their submissions. Correlation matrices should include Ns, means, SDs, and correlation matrices for all variables included in the analytical models (including control variables, dummy variables, interaction terms, transformed variables, etc.), and for all subgroups if applicable.

(2) Report coefficient estimates, SEs, sample sizes and exact *p*-values (no asterisks or cut-off levels) in all regression models (Bergh et al., in press; 2016).

The above requirements have been published by APA publication manual (American Psychological Association, 2010) and recently by *Strategic Management Journal* (SMJ, Bettis, Ethiraj, Gambardella, Helfat, & Mitchell, 2016). *MOR* new policies say the journal will "require authors to report coefficient estimates alongside exact *p*-values or standard errors" (Lewin et al., forthcoming). We would like to strongly recommend the requirement to replace "or" with "and". Only such full disclosure of data and statistics could enable verification of study's findings and future replication studies. New policies to enhance data and procedure transparency have been implemented at journals such as *Organizational Behavior and Human Decision Process*, *American Economic Review* and soon *MOR.* As reporting requirements evolve in different disciplines, there are more reporting requirements that need to be in place. We urge our academic community to continuously discuss and debate what reporting requirements we should follow.

2. How well are we trained in the use and interpretation of statistics?

The findings of missing statistical information prompts us to question the training of researchers to use, report and interpret statistics. They need to understand why the full reporting of statistical results is essential.

When Bedeian asked his graduate students a seemingly innocent question "What do you see when you look at a standard correlation matrix with its accompanying descriptive statistics?", the question was met with blank stares. He therefore introduced a 12-point checklist when reading correlation tables to identify the most basic aspects of a study's primary data, as these tables often can reveal "more than meets the eye" (Bedeian, 2014). The tests we adopt in this essay offer other insights into the value of the correlation matrix. We

encourage all PhD students, researchers, reviewers and editors to answer the question "what do we see when you look at a correlation matrix and regression report table?".

A similar innocent question is what does a *p*-value mean? NHST is a starting point for many analytic approaches ranging from regression analysis to multilevel analysis. However *p*-values have been widely misinterpreted among researchers as Bettis et al pointed out "it is incorrect to interpret *p* as the probability that the null hypothesis H0 is false, instead *p* is the probability that the sample value would be at least as large as the value actually observed if the null hypothesis is true" (Bettis et al., 2016, p259). Hence the use of cut-off level of *p*-values to support or reject hypotheses is inappropriate. SMJ has already implemented below policy:

"*SMJ will no longer accept papers for publication that report or refer to cut-off levels of statistical significance (p-values). In statistical studies, authors should report either standard errors or exact p-values (without asterisks) or both, and should interpret these values appropriately in the text. Rather than referring to specific cut-off points, the discussion could report confidence intervals, explain the standard errors and/or the probability of observing the results in the particular sample, and assess the implications for the research questions or hypotheses tested… SMJ will require in papers accepted for publication that authors explicitly discuss and interpret effect sizes of relevant estimated coefficients*" (Bettis et al., 2016, p 261).

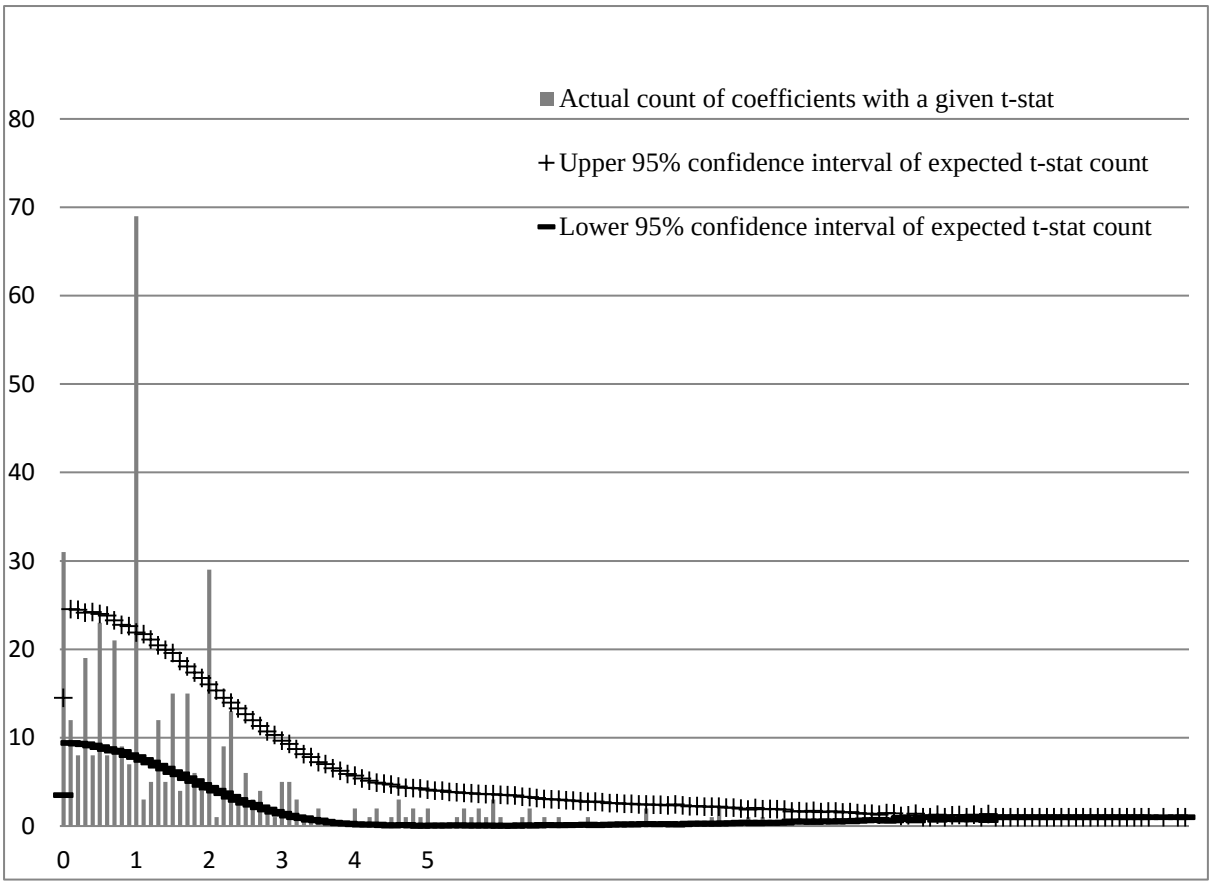
A full report of regression coefficients, SEs/t-values, and absolute *p*-values is not only precondition for verification, but also for providing a full interpretation of regression statistics. We believe discussion about fundamental questions in the use, reporting and interpretation of statistics will inform the training and development of current and future researchers and contribute toward increasing the verifiability and accuracy of reported empirical results. Such a conversation will substantiate the confidence that readers place in an article's conclusions, and ensure a more solid accumulative knowledge base.

Endnote

Innocent not knowing may lead to innocent not reporting. Not reporting leads to not being able to independently verify. Without verification, we lose the ability to confidently replicate research which can threaten the credibility of scientific knowledge. This essay reports a relatively high verifiability rate of *MOR* articles, but also highlights some patterns of reporting practices do not sufficiently facilitate verification. We call for more formalization of the role of verification within the review and evaluation process, specific attention to training researchers on the importance of data disclosure, and a slight change to the reporting requirements in all empirical *MOR* publications to facilitate independent verification. The journal's credibility and future leadership in management and organization research, particularly in shaping the development of indigenous research, depends on such steps.

FIGURE 1

Results of Test Three



REFERENCE

American Psychological Association. 2010. ***Publication manual of the American Psychological Association*** Washington, DC: Author.

Bakker, M., & Wicherts, J. 2011. The (mis)reporting of statistical results in psychology journals. ***Behavior Research Methods***, 43(3): 666-678.

Bedeian, A. G. 2014. "More Than Meets the Eye": A Guide to Interpreting the Descriptive Statistics and Correlation Matrices Reported in Management Research. ***Academy of Management Learning & Education***, 13(1): 121-135.

Bedeian, A. G., Taylor, S. G., & Miller, A. N. 2010. Management Science on the Credibility Bubble: Cardinal Sins and Various Misdemeanors. ***Academy of Management Learning & Education***, 9(4): 715-725.

Bergh, D., Sharp, B., & Li, M. in press, 2017. Tests for identifying "red flags" in empricial findings: Demonstration and recommendations for authors, reviewers and editors. ***Academy of Management Learning & Education,*** 16(1): 110-124.

Bergh, D., Sharp, B., Aguinis, H., & Li, M. in press. Is there a credibility crisis in strategic management research? Evidence on the reproducibility of study findings, ***Strategic Organization***

Berle, D., & Starcevic, V. 2007. Inconsistencies between reported test statistics and p-values in two psychiatry journals. ***International Journal of Methods in Psychiatric Research***, 16(4): 202-207.

Bettis, R. A., Ethiraj, S., Gambardella, A., Helfat, C., & Mitchell, W. 2016. Creating repeatable cumulative knowledge in strategic management. ***Strategic Management Journal***, 37(2): 257-261.

García-Berthou, E., & Alcaraz, C. 2004. Incongruence between test statistics and P values in medical papers. ***BMC Medical Research Methodology***, 4(1): 13.

Goldfarb, B., & King, A. A. 2016. Scientific apophenia in strategic management research: Significance tests & mistaken inference. ***Strategic Management Journal***, 37(1): 167-176.

John, L. K., Loewenstein, G., & Prelec, D. 2012. Measuring the Prevalence of Questionable Research Practices With Incentives for Truth Telling. ***Psychological Science***.

Levine, S. S. 2012. Walter R. Nord and Ann F. Connell: Rethinking the Knowledge Controversy in Organization Studies: A Generative Uncertainty Perspective. ***Administrative Science Quarterly***, 57(3): 537-540.

Lewin, A. Y., Chiu, C.-Y., Fey, C. F., Levine, S. S., McDermott, G., Murmann, J. P., & Tsang, E. forthcoming. The Critique of Empirical Social Science: New Policies at *Management and Organization Review*. ***Management and Organization Review***.

www.dictionary.com.

Zientek, L. R., & Thompson, B. 2009. Matrix summaries improve research reports: Secondary analyses using published literature. ***Educational Researcher***, 38: 343-352.